Social Cognitive and Affective Neuroscience, 2018, 933–943

doi: 10.1093/scan/nsy061
Advance Access Publication Date: 22 August 2018
Original article

Dissociating neural signatures of mental state retrodiction and classification based on facial expressions

Kathleen Kang,[1,3] Dana Schneider,[2] Stefan R. Schweinberger,[2] and Peter Mitchell[1]

[1]School of Psychology, the University of Nottingham, Nottingham, NG7 2RD, UK, [2]Institute of Psychology, Friedrich Schiller University of Jena, Jena, 07737, Germany and [3]Yong Loo Lin School of Medicine, National University of Singapore, 119228, Singapore

Correspondence should be addressed to Prof. Peter Mitchell, PhD, School of Psychology, University Park Campus, The University of Nottingham, Nottingham NG7 2RD, England. E-mail: peter.mitchell@nottingham.ac.uk.

Abstract

Posed facial expressions of actors have often been used as stimuli to induce mental state inferences, in order to investigate 'Theory of Mind' processes. However, such stimuli make it difficult to determine whether perceivers are using a basic or more elaborated mentalizing strategy. The current study used as stimuli covert recordings of target individuals who viewed various emotional expressions, which caused them to spontaneously mimic these expressions. Perceivers subsequently judged these subtle emotional expressions of the targets: in one condition ('classification') participants were instructed to classify the target's expression (i.e. match it to a sample) and in another condition ('retrodicting') participants were instructed to retrodict (i.e. infer which emotional expression the target was viewing). When instructed to classify, participants showed more prevalent activations in event-related brain potentials (ERPs) at earlier and mid-latency ERP components N170, P200 and P300–600. By contrast, when instructed to retrodict participants showed enhanced late frontal and fronto-temporal ERPs (N800–1000), with more sustained activity over the right than the left hemisphere. These findings reveal different cortical processes involved when retrodicting about a facial expression compared to merely classifying it, despite comparable performance on the behavioral task.

Key words: Theory of mind; facial expressions; social cognition; event-related potentials; retrodictive mentalizing

Introduction

A widely investigated mentalistic inference is prediction of a target's behavior from knowledge of the target's state of belief on, say, the location of a desired object (Premack & Woodruff, 1978; Leslie, 1987; Perner, 1991; Fodor, 1992; Apperly, 2013). But a common form of mentalistic inference in real life is retrodiction of an antecedent event to explain the target's ongoing behavior (Gallese and Goldman, 1998). Here a common case is facial expression. When seeing a facial expression (i.e. the target's behavior), we might be inclined to determine the cause of that behavior. In such an event, a mental operation will be set in motion to make the relevant retrodictive inference (Kang *et al*., 2017). If observers (henceforth 'perceivers') are not inclined to make a retrodictive inference, they will not venture beyond merely classifying (e.g. labeling) the expression. Mentalizing, including retrodiction, does not always happen automatically (Apperly *et al*., 2006), and in the absence of a cue, perhaps by

Received: 5 May 2018; **Revised:** 26 June 2018; **Accepted:** 24 July 2018

default, perceivers will merely classify the expressions. Here we asked whether mental state retrodiction and classification are distinct processes associated with dissociable neural signatures, using electroencephalography (EEG).

Two advantages of investigating participants' ability to make a retrodictive inference are that (1) the associated method is well suited to satisfying the 'truth condition' (see West & Kenny, 2011) which states that participants' responses should be measured against an objective fact, and (2) using natural, spontaneous and subtle stimuli that are closer to what is experienced in real-life encounters. While target behaviors can have various causes (Pillai *et al*., 2012, 2014; Cassidy *et al*., 2014, 2015; Sheppard *et al*., 2016; Teoh *et al*., 2017), one such cause with respect to facial expressions is the natural mirroring that occurs when people meet (Dimberg *et al*., 2000; Sato & Yoshikawa, 2007).

To capture this phenomenon experimentally, Kang *et al*. (2017) showed a succession of emotionally expressive faces on a laptop to target participants as they were unknowingly recorded by the laptop's integral camera. Images of the targets spontaneously and subtly mirroring the displayed expressions were then shown to another group of participants, perceivers, who were tasked with inferring which expression the target had been observing. The truth condition was satisfied such that the perceiver's inference can be compared and indeed be validated against an objective fact (i.e. which expression the target was actually seeing). This task contrasts with methods which employ standardized facial expressions, where the target's expression is posed (i.e. does not have any natural cause) and where, strictly speaking, it is invalid to ask the perceiver to make a mentalistic inference as to the underlying cause of the target's behavior.

In the research presented here, we divided perceivers in two groups. One was instructed to classify the target expressions while the other was instructed to infer the cause of the target's expression. The latter uniquely cues perceivers to make a retrodictive inference, thus to engage in a deeper level of social-cognitive processing (Apperly *et al*., 2006). It would not be appropriate to test each perceiver under both conditions as it would be unreasonable to expect them to 'unlearn' the first received instruction. It is assumed that the wording of the question and not the inherent properties of the target faces, cues perceivers to classify or retrodict; and we assume they might not retrodict without being cued to do so (Apperly *et al*., 2006). Nevertheless, it is important to use target faces that convey a signal such that the task is soluble. If the task was insoluble, then there would be no evidence that perceivers were engaging any kind of social cognitive processing.

From the face processing literature, it has been suggested that emotional expressions are processed temporally in a few steps. The first entails structural encoding, which evokes the N170 component (Bentin *et al*., 1996; Maurer *et al*., 2008; Kloth *et al*., 2010). Many assume that the N170 is not modulated by the emotional signal in faces (Eimer & Holmes, 2002; Ashley *et al*., 2004). However, Smith (2012) suggested that the N170 may also reflect an emotional categorization process (Batty & Taylor, 2003; Blau *et al*., 2007). Following on from the N170, an emotional analysis might also be reflected in the P200 component (Paulmann & Pell, 2009). Later, a detailed perceptual analysis of affective significance is shown in mid-latency components (approximately 300 ms; Cuthbert *et al*., 2000) and even later, cognitive evaluation, which involves stored knowledge about the emotional expressions, is demonstrated by the N400 component (Eimer, 2000; Adolphs, 2002).

In terms of neural correlates of mentalizing, fMRI studies have consistent activation in three areas, namely the medial prefrontal cortex (mPFC) and the left and the right temporo-parietal junction (TPJ-L and TPJ-R, respectively; Schurz *et al*., 2014). The TPJ-R has consistently shown an increase in activation during false belief tasks (Vogeley *et al*., 2001; Saxe and Kanwisher, 2003) and thus is said to be specifically sensitive to processing belief information (Aichhorn, 2006) and thinking about thoughts (Saxe, 2007, 2009; Young *et al*., 2010). On the other hand, the TPJ-L has been found to be equally responsible for processing false beliefs and false statements (Aichhorn, 2006), suggesting its role in helping people think about the idea of a representation (Saxe, 2007). Activation in these TPJ regions were also demonstrated when participants were asked to infer others' inner states from the eyes region alone ('Mind in the Eyes' task—Baron-Cohen *et al*., 2001; Schurz *et al*., 2014). Additionally, the mPFC was also activated during social and emotional information processing (Aichhorn *et al*., 2006). Frontal and parietal slow wave activities were prominent during explicit mentalizing (Sabbagh & Taylor, 2000; Liu *et al*., 2004, 2009; Geangu *et al*., 2012). Activity over the frontal sites and enhanced positivity in the late positive component differentiated between reasoning about beliefs and reality (Sabbagh and Taylor, 2000) and between false belief and true belief (Meinhardt *et al*., 2011). Additionally, McCleery *et al*. (2011) suggested that perspective taking was reflected in bilateral temporal-parietal regions and lateral frontal regions. Late slow wave activity over the frontal and right posterior areas of the scalp (Liu *et al*., 2009) was shown after 500–1000 ms from the onset of the experimental stimulus for belief judgments. Although the ERP components differ slightly depending on the type of mentalizing task, it has been consistently shown that a more elaborated form of mentalizing is reflected in late slow wave activity over the frontal and parietal regions. Only one ERP study reported in the literature looked at inferring mental states from facial expressions (i.e. Sabbagh *et al*., 2004). The authors identified a stronger N270-400 over the right inferior frontal and right anterior temporal regions when perceivers were making mental state judgments, as compared to gender judgments. These findings allow several possible explanations: one explanation supposes that when perceivers were asked to guess the mental state of targets they were genuinely and uniquely inferring the target's mental states through a more elaborated form of mentalizing. An alternative explanation supposes that the experimental condition merely involved a basic form of mentalizing (i.e. classification). However, performance in inferring the mental state of targets was compared against a gender classification condition, a classification dimension that arguably is rather unrelated to internal mental states. This makes these findings of ERP differences between mentalizing and classifying less remarkable. It is not known whether the ERP findings in the mental state condition reflect a basic or elaborated form of mentalizing. In a related study, Sessa *et al*. (2014) discovered that facial expressions of pain selectively modulated activity at 110–360 ms over fronto-central and centro-parietal regions while painful contexts selectively modulated activity at 400–840 ms over fronto-central and centro-parietal regions, hence demonstrating dissociation between the perceptual and cognitive component of social cognition.

Thus, in the current study we included experimental conditions to disentangle ERP activity involved in basic mentalizing strategies, compared with a more elaborated mentalizing strategy, namely, 'retrodiction'. We expected to find broad processing differences between retrodiction and classifying that transcend any specific differences in processing different types of emotional stimuli. If retrodiction genuinely involves a more

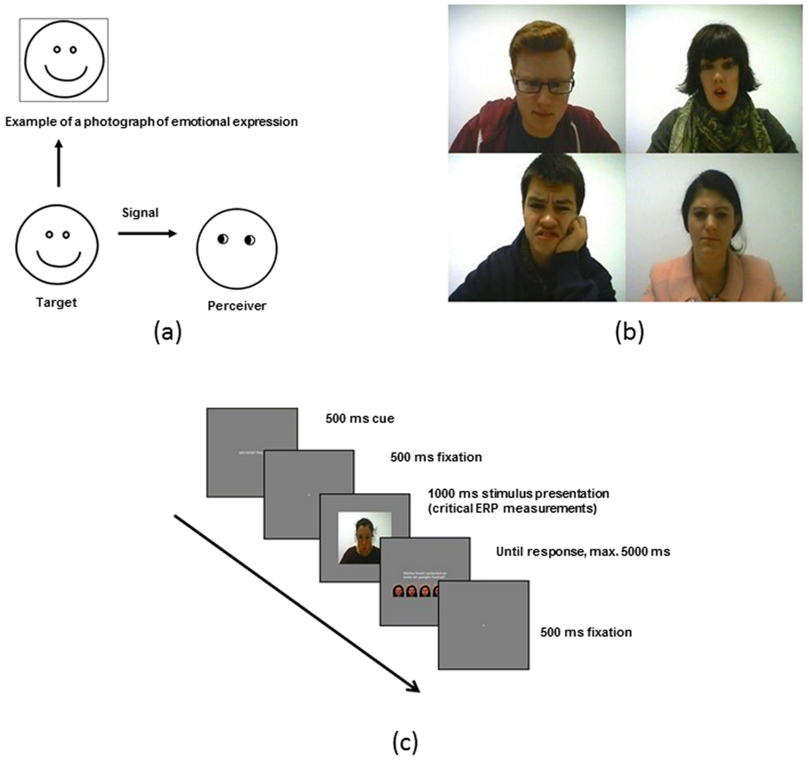

Fig. 1. A schematic diagram illustrating the two stages of the experiment. The target looked at the photograph of an expression while his/her own expression was covertly recorded ('Target Phase'). The perceiver then watched the recording of the target and guessed what photograph the target was looking at ('Perceiver Phase'); (b) example of peak expressions used as stimuli—(from left to right) happiness, surprise, disgust and anger; (c) experimental paradigm for both classifying and retrodicting perceivers in the 'Perceiver Phase'. Participants saw a cue word for 500 ms, followed by a fixation cross for 500 ms. Then, the target's expression was displayed for 1000 ms. After that, participants had to make a response on the judgment scale (depicting Ekman-type expressions of happy, surprised, anger, disgusted and neutral) which was displayed for 5000 ms. Finally, a 500 ms fixation cross appeared.

elaborated form of mentalizing while classifying does not, then one might expect a more prominent N270-400 in the retrodiction task as compared to the classification task. According to Sabbagh *et al.*'s (2004) findings, we would expect greater activity at N270-400 in perceivers who were cued to retrodict; and we expect these perceivers will show stronger activation over later components at the frontal and temporo-parietal electrodes (particularly, right temporo-parietal electrodes). Since classification perceivers were not explicitly cued to make a retrodictive inference, we would expect activation similar to facial processing components, specifically earlier components (N170, P200) and mid-latency components.

Material and methods

This study was divided into two stages: the (A) 'Target Phase' and the (B) 'Perceiver Phase'. In (A) targets were asked to look at photographs of facial expressions while being video recorded covertly (see Supplementary Materials for an extensive description of the target phase and stimuli). Their videos were then shown to perceivers who, in the retrodiction condition, had to guess which expressions the targets were looking at, satisfying the 'truth condition' (West and Kenny, 2011). Figure 1A illustrates the relationship between the two stages of the study.

Perceiver phase

Thirty-two participants (15 males, 17 females; mean age, 25.75 years, s.d. = 4.74 years) were recruited as perceivers from the Friedrich Schiller University of Jena, Germany. Before perceivers took part in the study, they confirmed that they understood what they were being asked to do and provided written consent. We used a between-subjects design to eliminate the possibility of carry-over effects. Parity between groups was ensured by unbiased allocation of participants to conditions. To avoid any confound associated with experimenter effects, the same researcher was responsible for setting up the experiment across conditions, including fitting the cap to participants. During the experimental procedure (see Figure 1C) perceivers first saw a 'Get Ready' screen for 500 ms, followed by a fixation cross displayed for a further 500 ms. After that, they saw a photograph of the target's facial expression displayed for 1000 ms (one of the four expressions and filler trials). Subsequently, perceivers saw a photorealistic face-scale depicting Ekman-type expressions of happiness, surprised, anger, disgusted and neutral for a maximum of 5000 ms. This emotional face judgment scale displayed the selection of basic emotions seen by targets during the 'Target Phase'. The gender of the scale was matched to the target's gender (i.e. a male target face was followed by a male emotional judgment scale) to ensure perceivers in the classification task were not unintentionally cued about the mental state of targets. Half of the perceivers were randomly assigned to the classification task and were asked, 'Which picture best represents the displayed expression?' Perceivers in this task were told they would see several facial expressions and were required to choose the Ekman face that best represents the facial expressions they saw. The remaining perceivers were assigned to the retrodiction task and were asked, 'Which picture did the person look at?' Perceivers in this task

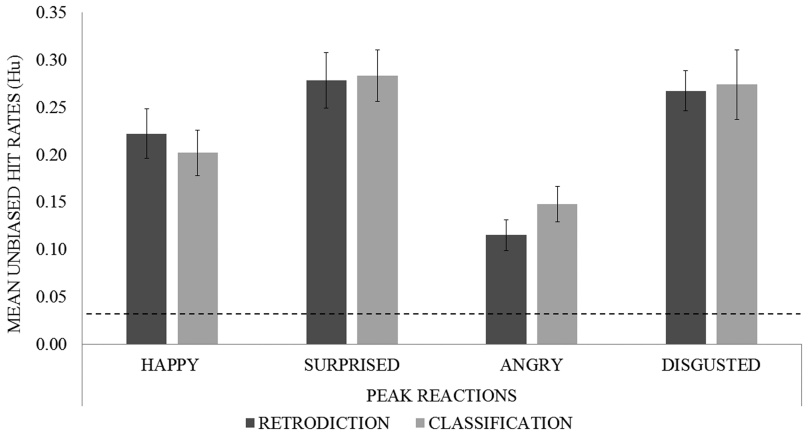

Fig. 2. Mean unbiased hit rates for all viewed target peak expressions (i.e. happy, surprise, angry and disgust) for both perceiver groups. Error bars represent the SEM. The dotted line represents chance level performance.

were told that they would see several facial expressions, which were the target's reactions to a face they had seen. After the perceiver made a response, another fixation cross was displayed for 500 ms. Perceivers in both groups were matched in terms of age, gender, handedness and scores on the Autism Spectrum Quotient (Baron-Cohen *et al.*, 2001).

Results

Behavioral results

A confusion matrix (Table 1 in the Supplementary Materials) shows that perceivers tended to judge targets as showing a neutral expression. Thus, Wagner's (1993) unbiased hit rate (Hu) was calculated (a method used extensively in emotion recognition research, see Tcherkassof *et al.*, 2007; Pell *et al.*, 2009; Orgeta, 2010) to ensure that judgmental accuracy was not influenced by that bias. Hu expresses accuracy as a proportion of both response frequency and stimulus frequency, ranging from 0 (responses never correspond with the respective stimulus category) to +1 (responses frequently corresponded with the respective stimulus category). One-sample t-tests were conducted to determine if perceivers were systematically able to recognize target expressions. Bonferroni correction was applied ($P = 0.0125$) to minimize false positives. Perceivers in both retrodiction and classification tasks were able to detect all facial expressions reliably (i.e. happy ($t(15) = 6.93$, $P < 0.001$; $t(15) = 6.87$, $P < 0.001$), surprised ($t(15) = 8.02$, $P < 0.001$; $t(15) = 8.96$, $P < 0.001$), angry ($t(15) = 4.30$, $P = 0.001$; $t(15) = 5.28$, $P < 0.001$) and disgusted ($t(15) = 10.14$, $P < 0.001$; $t(15) = 6.02$, $P < 0.001$), respectively).

A 2 (task: classification and retrodiction) × 4 (target peak expression: happiness, surprised, anger and disgusted) mixed analysis of variance (ANOVA) was conducted. Where appropriate, Epsilon corrections for heterogeneity of covariances were performed. A significant main effect of target expression, $F(3, 90) = 22.02$, $P < 0.001$, $\eta_p^2 = 0.42$, $\epsilon = 0.90$, indicated perceivers were systematically more accurate for happy than angry, $P < 0.001$; surprised than happy, $P = 0.022$, and angry, $P < 0.001$; and disgusted than angry, $P < 0001$. Note, there was no main or interaction effect involving the factor task, $Fs < 1$, $Ps > 0.10$. See Figure 2. Further *post hoc* tests were therefore not conducted.

A 2 (task: classification and retrodiction) × 4 (target peak expression: happy, surprised, angry and disgusted) mixed ANOVA was conducted on perceiver response times. There was a significant main effect of target expression, $F(3, 90) = 6.64$, $P = 0.001$, $\eta_p^2 = 0.18$, $\epsilon = 0.92$, and pairwise comparisons revealed that perceivers responded faster to happy as compared to angry, $P = 0.006$, and disgust target expressions, $P = 0.013$. No other effects were significant, including effects associated with the perceiver task, $Fs < 1.6$, $Ps > 0.10$.

Overall, the behavioral findings did not show any differences between the classification and retrodiction task, suggesting the two tasks were equivalent and that they had been allocated without bias.

Electrophysiological data

A five-way mixed ANOVA with perceiver task (i.e. classification and retrodiction) as a between-subjects factor and hemisphere, site and target expression as within-subjects factors was conducted for the N170, P200, N270-400, N600-800 and N800-1000. We measured mean amplitudes to quantify these components on the assumption that these are particularly appropriate for later ERP components (Luck, 2005), which are the focus of interest for this study.

Electroencephalographic data were recorded with a 64-channel BioSemi™ (BioSemi, Amsterdam, Netherlands) Active Two-System. The positioning of the electrodes was as follows: FP1, FT9, AF3, F1, F3, F5, F7, FT7, TP9, FC3, FC1, C1, C3, C5, T7, TP7, PO9, CP3, CP1, P1, P3, O9, P7, P9, PO7, PO3, O1, Iz, Oz, POz, Pz, CPz, FPz, FP2, FT10, AF4, AFz, Fz, F2, F4, F6, F8, FT8, TP10, FC4, FC2, FCz, Cz, C2, C4, C6, T8, TP8, PO10, CP4, CP2, P2, P4, O10, P8, P10, PO8, PO4 and O2. As for reference electrodes, the BioSemi™ system utilizes two additional electrodes, Common Mode Sense (CMS) and Driven Right Leg (DRL), which form a feedback loop that drives the average potential of participants as close as possible to the Analogue-to-Digital Converters (ADC) reference in the AD box (for further information, see http://www.biosemi.com/faq/cms&drl.htm). Horizontal electro-oculogram (EOG) was recorded by placing two electrodes on the outer canthi of both eyes. Meanwhile, vertical EOG was recorded with two electrodes placed above and below the left eye. The EEG data were amplified using a BioSemi™ amplifier and digitally recorded with ActiView™ (version 6.0.5) using a sampling rate of 512 Hz and online filtering (DC to 120 Hz, low-pass). The data were then pre-processed in BESA™ (Brain Electromagnetic Source Analysis, version 5.1.8, Gräfelfing, Germany). Offline, ocular artifacts were corrected using a multiple source approach implemented in BESA (Berg & Scherg, 1994). Note that eye movement correction applied to

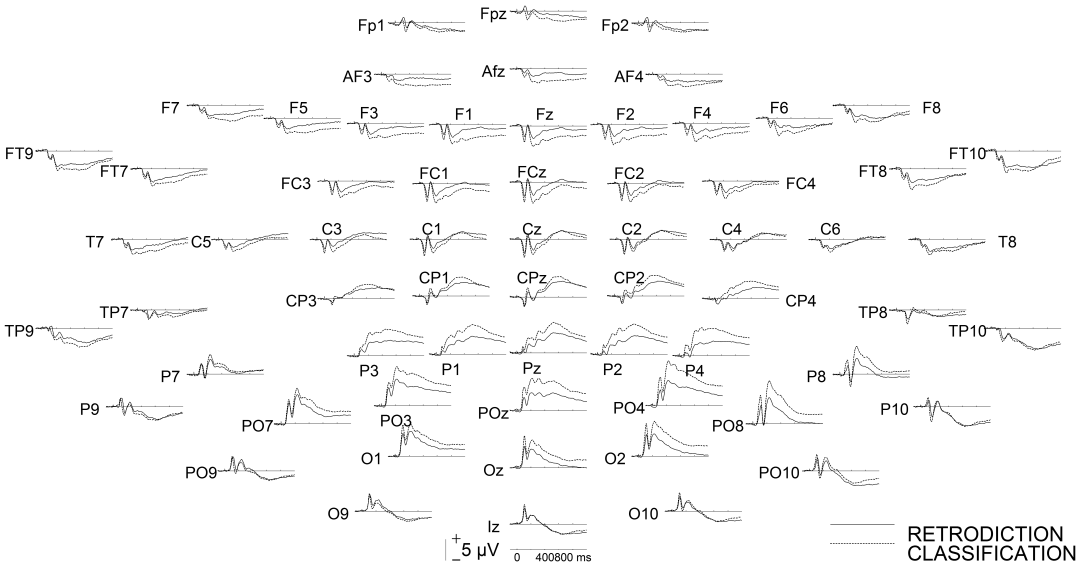

Fig. 3. N170 early occipito-temporal ERPs across hemispheres as a function of target expression for the classification group only. Note the larger amplitudes over the right compared to the left hemisphere. In the retrodiction group, no hemisphere by target expression interaction was found.

activity recorded at EEG electrodes excludes the possibility that observed EEG differences between experimental conditions were influenced by differential eye movements. Non-ocular artifacts were excluded using BESA^{TM}'s artefact rejection tool (amplitude threshold of 100 μV amplitude and 75 μV gradient). Offline, EEG data were recalculated to the average reference and were low-pass filtered at 40 Hz using a zero-phase shift digital filter. After pre-processing, between 90% and 95% of all trials were accepted and analyzed, and these numbers were similar across experimental conditions. On average, 93.7% and 95.0% of all trials were analyzed in the retrodiction and classification tasks, respectively. Any differences between numbers of accepted trials were negligible across different emotions, with averages of 93.8% for happy trials, 95.0% for surprise trials, 94.0% for angry trials and 94.5% for disgust trials, regardless of task.

Epochs with a 200 ms pre-stimulus baseline and duration of 1200 ms were generated. Trials for each electrode as well as each experimental condition/task were averaged separately. ERPs were quantified by measuring mean amplitudes for the following components and in the following time windows: N170 (140—170 ms), P200 (220—250 ms), N270—400 (270—400 ms), P300—600 (300—600 ms), N300—600 (300—600 ms) and N800—1000 (800—1000 ms). These time windows and the measurement electrodes for these ERPs were chosen based on previous research and on visual inspection of the grand averages. For the early components with a clear peak (N170, P200), we additionally ensured that mean peak latencies as visually inspected were stable across conditions in the grand averages (cf. Figure 3 for examples). Small time windows were used for mean amplitude measurements that were centered on the mean latencies of the peaks. Specifically, the N170 was measured at P9/P10, P7/P8, PO7/PO8 and PO9/PO10. The P200 was measured at PO3/PO4 and PO7/PO8. Later components exhibited less clear peaks, potentially due to trial-to-trial variability in the timing of neurocognitive processes, which is known to be more prevalent for later ERPs. Mean amplitude measures are relatively robust to trial-to-trial variability in component latency (Luck, 2005) and therefore were also used for later components. The N270-400 was quantified at FT7/FT8 and FT9/FT10. The parietal positivity of the P300-600 was quantified at P3, P1, Pz, P2 and P4. The fronto-temporal N600-800 was measured at FT7/FT8, FT9/FT10, and the N800-1000 was quantified at FT7/FT8, FT9/FT10, F5/F6, F7/F8, TP7/TP8 and TP9/TP10. Note that consecutive odd and even numbers denote homologous sites over the left and right hemisphere, respectively.

For ease of interpretation and readability, only significant main and interaction effects of condition/task are reported here. Whenever there were higher-order interactions involving perceiver condition/task as a factor, further analyses were conducted for the two tasks independently. Pairwise comparisons using multiple *t*-tests were conducted with Bonferroni correction. For an illustration of overall ERP differences between the retrodiction and classification tasks, refer to Figures 4 and 5, which depict waveform data and topographical voltage maps, respectively (see Supplementary Materials for more extensive ERP data analyses).

Hypothesis 1. Did the classifying task elicit a stronger activation over the earlier and mid-latency components, as compared to the retrodiction task?

N170 (P7/P8, P9/P10, PO7/PO8 and PO9/PO10). An analysis of the peak expressions for the N170 component revealed a significant Hemisphere × Target Expression × Task interaction, $F(3,90)=3.75$, $P = 0.014$, $\eta_p^2 = 0.10$, $\epsilon = 0.96$. In the classification task, there was a significant interaction between 'Hemisphere' and 'Target Expression', $F(3,45) = 6.45$, $P = 0.002$, $\eta_p^2 = 0.30$, $\epsilon = 0.89$. Over the left hemisphere, a main effect of 'Target Expression', $F(3,45) = 4.08$, $P = 0.012$, $\eta_p^2 = 0.21$, $\epsilon = 1.00$, revealed stronger negative activation for angry compared to surprised expressions, $P = 0.027$. Over the right hemisphere, there was a main effect of 'Target Expression', $F(3,45) = 3.57$, $P = 0.025$, $\eta_p^2 = 0.19$, $\epsilon = 0.91$, but pairwise comparisons did not identify any significant differences. Overall a main effect of hemisphere revealed stronger negative activation over the right compared to the left hemisphere for happy, $t(15) = 3.53$, $P = 0.003$; surprised, $t(15) = 2.14$, $P = 0.049$; angry, $t(15) = 3.86$, $P = 0.002$; and disgusted, $t(15) = 2.14$, $P = 0.014$, cf. Figure 3. By contrast, no main or interaction effects were found for the retrodiction task, $Fs < 2$, $Ps > 0.1$.

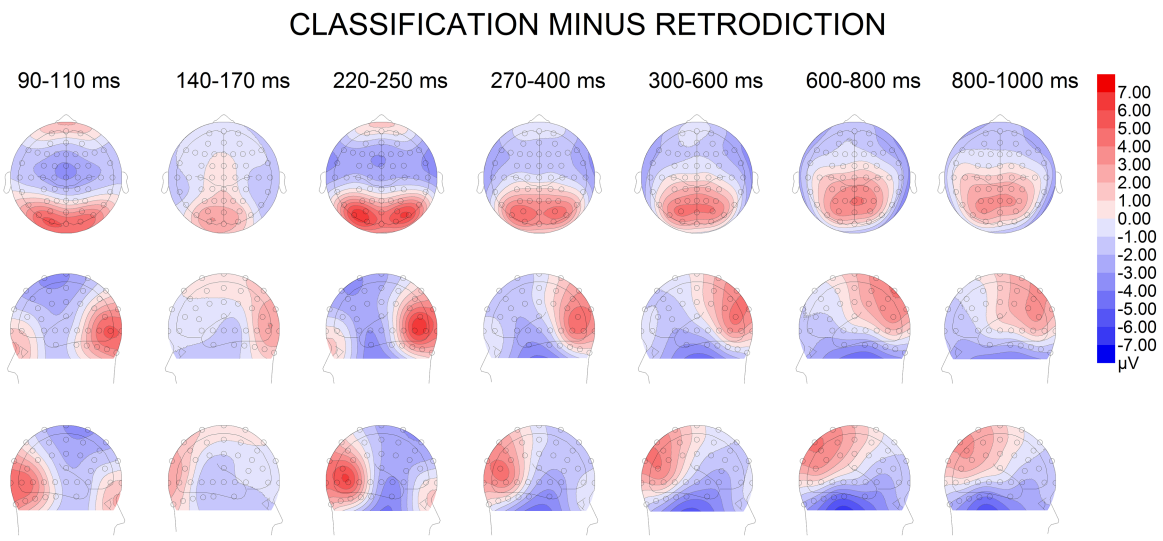

Fig. 4. ERP waveforms for the retrodiction and classification groups across all 64 channels, averaged across all other experimental conditions.

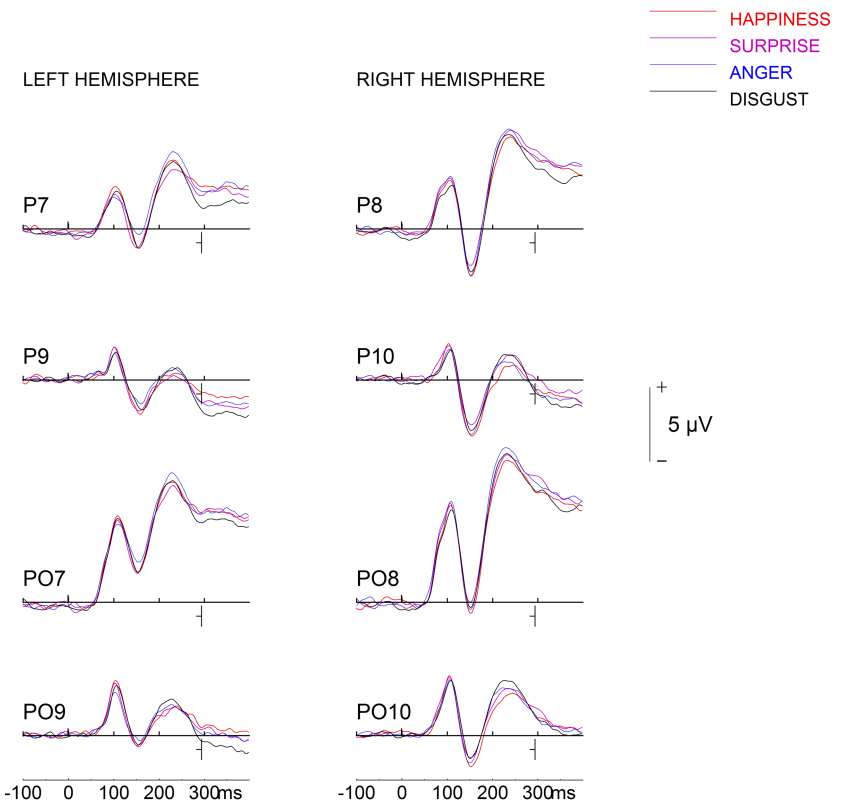

Fig. 5. Voltage maps depicting the topographical differences between groups, for all time segments analyzed, and collapsed across the other experimental conditions. The top row shows a top view map using a 110-degree equidistant projection. Middle and bottom rows show a left and right view, respectively, both using a 90-degree equidistant projection. All maps were obtained using spherical spline interpolation. Electrode positions are also shown. Negativity is blue.

The N170 analysis also revealed a significant Site × Task interaction, $F(3,90) = 3.32$, $P = 0.045$, $\eta_p^2 = 0.10$, $\epsilon = 0.64$. In the classification task, there was stronger negative activation over PO9/PO10 than P9/P10, $P = 0.001$; PO7/PO8 than PO9/PO10, $P < 0.001$; and PO7/PO8 as compared to P7/P8, $P < 0.001$. In the retrodiction task, there was stronger negative activation over PO9/PO10 than P9/P10, $P = 0.001$; PO7/PO8 as compared to P9/P10, $P = 0.009$; and PO7/PO8 than P7/P8, $P < 0.001$.

P200 (PO3/PO4 and PO7/PO8). There was a significant main effect of perceiver task, $F(1,30) = 9.40$, $P = 0.005$, $\eta_p^2 = 0.24$, reflecting a significantly larger P200 amplitude for classification compared to the retrodiction task (cf. Figures 4 and 5). All other effects and interactions were not statistically significant, $Fs < 2.36$, $Ps > 0.10$.

P300-600 (P3, P1, Pz, P2 and P4). There was a significant main effect of perceiver task, $F(1,30) = 17.67$, $P < 0.001$, $\eta_p^2 = 0.37$, in

that there was significantly larger parietal positive activation for perceivers in the classification task than in the retrodiction task. All other effects and interactions were not statistically significant, $Fs < 1.15$, $Ps > 0.20$.

P300-600 (PO7, PO3, POz, PO8 and PO4). There was a significant main effect of perceiver task, $F(1,30) = 8.75$, $P = 0.006$, $\eta_p^2 = 0.23$, with stronger positive activation for the classifying perceivers as compared to the retrodiction perceivers. All other effects and interactions were not statistically significant, $Fs < 1.43$, $Ps > 0.20$.

Hypothesis 2. Did the retrodiction task elicit more negativity (or less positivity) for N270-400 in comparison to the classifying task?

N270-400 (PO7/PO8, PO3/PO4 and P7/P8). There was a significant 'Site' × 'Task' interaction, $F(2,60) = 4.82$, $P = 0.014$, $\eta_p^2 = 0.14$, $\epsilon = 0.91$. There was also a significant main effect of 'Task', $F(1,30) = 6.57$, $P = 0.016$, $\eta_p^2 = 0.18$, in which there was more negativity (less positivity) for the retrodiction than the classification task. In the retrodiction task, there was a significant main effect of 'Site', $F(2,30) = 29.55$, $P < 0.001$, $\eta_p^2 = 0.66$, $\epsilon = 0.99$. There was stronger positivity over PO7/PO8 than P7/P8, $P < 0.001$, and PO3/PO4 as compared to P7/P8, $P < 0.001$. In the classification task, there was also a significant main effect of 'Site', $F(2,30) = 54.04$, $P < 0.001$, $\eta_p^2 = 0.78$, $\epsilon = 0.88$, with stronger positivity over PO7/PO8 than P7/P8, $P < 0.001$; PO3/PO4 than PO7/PO8, $P = 0.023$; and PO3/PO4 as compared to P7/P8, $P < 0.001$. As for task differences, there was more negativity (less positivity) for retrodiction perceivers than classification perceivers over PO3/PO4, $t(30) = 43.36$, $P = 0.002$. All other effects and interactions were not statistically significant, $Fs < 3.24$, $Ps > 0.08$.

Hypothesis 3. Did the retrodiction task elicit a stronger late slow wave activity over the frontal, fronto-temporal and temporo-parietal electrodes as compared to the classifying task?

N800-1000 (F7/F8 and F5/F6). There was a significant Hemisphere × Task interaction, $F(1,30) = 5.36$, $P = 0.028$, $\eta_p^2 = 0.15$ (cf. Figures 6 and 7). The classification task showed significantly stronger frontal negative activation over the left hemisphere, as compared to the retrodiction task, $t(30) = 2.44$, $P = 0.021$. By contrast, there were no significant differences between tasks over the right hemisphere. All other effects and interactions were not statistically significant, $Fs < 3.07$, $Ps > 0.09$.

N800-1000 (FT9/FT7 and FT10/FT8). There was a significant Hemisphere × Task interaction, $F(1,30) = 6.023$, $P = 0.02$, $\eta_p^2 = 0.17$. In the retrodiction task, there was stronger negative activation over the right hemisphere than the left hemisphere, $t(15) = 3.10$, $P = 0.007$. This difference was not seen in the classification task (see Figure 6, right half). All other effects and interactions were not statistically significant, $Fs < 3.13$, $Ps > 0.09$.

Discussion

The present study is the first to identify differences in neural responses to the very same faces when perceivers are asked to classify a facial expression (that might be an automatic process) or when perceivers are asked to determine what caused a facial expression (a retrodictive inference process activated on cue). Perceivers were accurate to a degree in making judgments—either in classifying or retrodicting—demonstrating that perceivers were engaging a social-cognitive process. The difference in neural signature across both tasks suggests that different processes were engaged depending on the question perceivers were asked. In general, the classification task elicited more prevalent ERP positivity compared to the retrodiction task at earlier ERP components: N170, P200 and P300-600. At late components (N800-1000), brain activity was driven by more sustained negativity over right than left fronto-temporal electrodes for the retrodiction task only. By contrast, for the classification task, more sustained negativity was found over left than right frontal electrodes. Overall, these results demonstrate dissociable cortical processes when one engages in retrodictive mentalizing from a facial expression compared to merely classifying it.

Specifically, in relation to the differences between cortical activity across the classification and retrodiction tasks, our results showed that activity of the N170 was modulated by different expressions only for classification. This finding is consistent with a number of previous studies, which found a modulation of the N170 according to different expressions (Blau *et al.*, 2007; Krombholz *et al.*, 2007; but see also Eimer and Holmes, 2002). Although Smith (2012) claims that the N170 is enhanced when an expression is categorized as emotional, the N170 was modulated according to the different expressions suggesting that different structural encoding processes were employed for different expressions. However, this was only the case when people were classifying a facial expression. Subsequent to the N170, we observed a considerably larger P200 response, which is involved in emotional decoding (Paulmann and Pell, 2009) for classification than retrodiction. This finding also supports Sessa *et al.*'s (2014) findings that the P200 is involved in the perceptual component of social cognition. Thus, it seems classification involves perceptual processing to a greater extent than retrodiction. Larger positivity for the P300-600 over parietal electrodes for the classification task supports previous research which showed that parietal ERPs in this time range are linked to emotional face processing, which in turn involves access to semantic memory (Eimer, 2000). The present larger P200 and subsequent components in the classification task may be tentatively related to recent findings in the context of socio-emotional language feedback. Specifically, Schindler and Kissler (2016) showed a strong and sustained amplification of ERP responses with an onset in the P200 time range when people saw emotional adjectives they interpreted as directed toward themselves only if they believed these originated from another human (as opposed to a machine). Arguably, classifying perceivers may have a tendency to perceive these emotions as being directed at themselves (cf. Adams and Kleck, 2003), whereas retrodicting perceivers might focus on the perspective of the other person and what caused them to make any given expression.

As for cortical activity in the retrodiction task, our findings did not replicate those reported by Sabbagh *et al.* (2004). Rather, we found a significant main effect of task (classifying and retrodiction) over parietal and occipito-temporal electrodes, with larger positivity in the classification task. These discrepancies between the present results and those by Sabbagh *et al.* (2004) may well be due to task differences, since their study contrasted mentalizing with a gender classification task whereas the present study used an expression classification task as the control condition. Arguably, perceptually classifying a facial expression only involves basic mentalizing processes, which are automatic. Retrodiction, on the other hand, may involve elaborated, controlled and flexible mental state representations (Apperly and Butterfill, 2009).

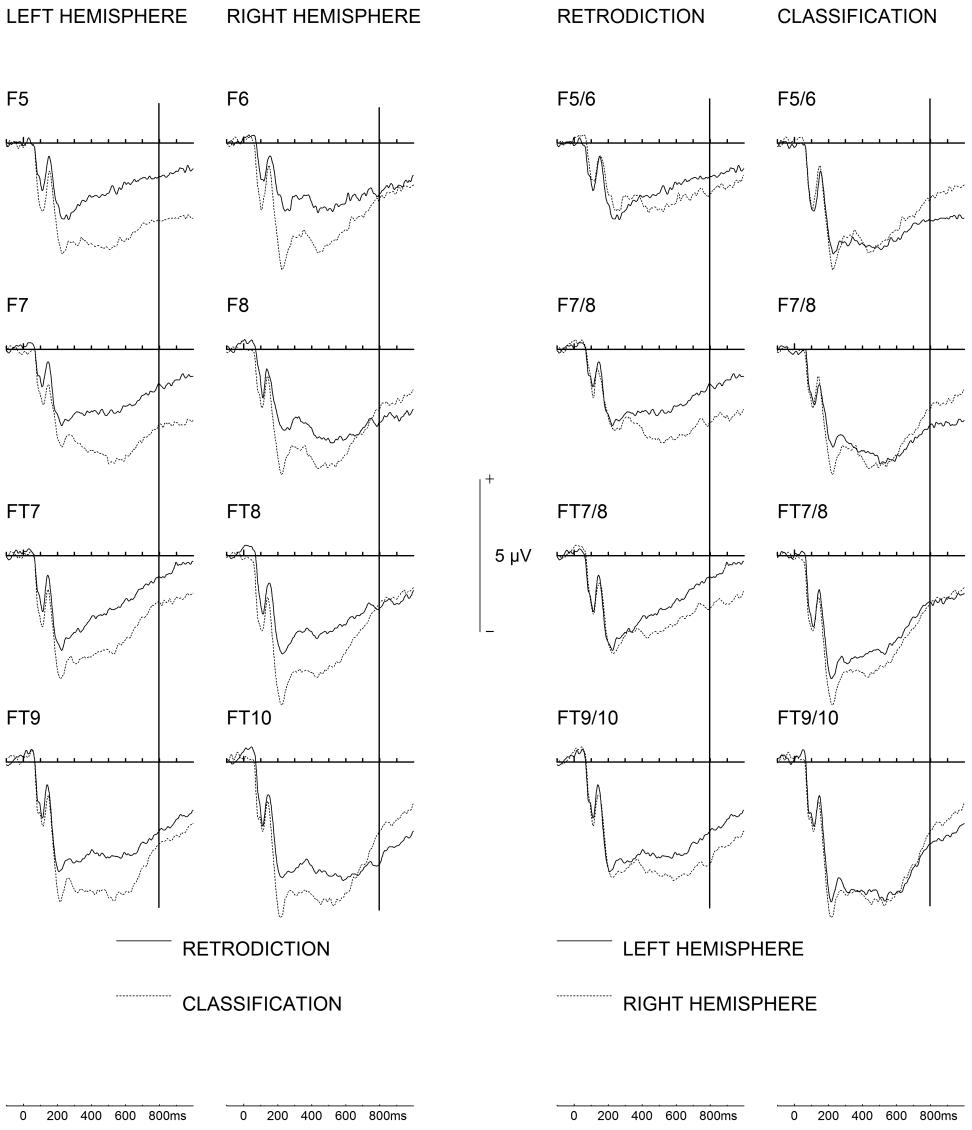

Fig. 6. Late frontal and fronto-temporal ERP asymmetries (800–1000 ms) for both groups. Left half of the figure: note that ERP negativity over the left hemisphere for the retrodiction group returns toward baseline whereas a distinct sustained activity over the right hemisphere can be observed. For the classification group, negativity over the right hemisphere returns toward baseline whereas a sustained negativity is seen over the left hemisphere. Right half of the figure: the same data plotted with ERPs from homologue electrodes over both hemispheres superimposed, such that hemispheric asymmetries become more readily visible. A vertical line is plotted at 800 ms.

At much later components between 800 and 1000 ms, activity in the classification task returned to baseline while there was sustained lateralized activity in the retrodiction task, over the right fronto-temporal and frontal electrodes. At the same time, late slow wave activity was more prominent for classification over the left hemisphere whereas activity for retrodiction returned to baseline, as indicated by a significant task difference over the left frontal electrodes. This sustained activity over the later components partially supports past studies that showed that frontal, parietal and central late slow wave activity is associated with mentalizing (Sabbagh and Taylor, 2000; Liu *et al.*, 2004; Meinhardt *et al.*, 2011; Geangu *et al.*, 2012; McCleery *et al.*, 2011). Moreover, the involvement of slow wave activity for the retrodiction task supports Sessa *et al.*'s (2014) suggestion that the cognitive component and higher-level processes of social cognition modulates activity of later ERP components. It should be noted that the present study was not designed to perform source analyses on the present scalp-recorded EEG data. In addition to the known problems with EEG source analyses, localization of the N800-1000 activity was unfeasible because of low signal amplitude and the possibility of simultaneous activity of multiple brain generators. However, our observation of sustained activity over the right hemisphere seems to be supported by previous imaging studies which emphasize the involvement of the right hemisphere in mentalizing, including the TPJ-R and right fronto-temporal regions (Schurz *et al.*, 2014). Other imaging studies implicate the role of the medial prefrontal cortex and right hemispheric involvement of other frontal areas including the middle frontal gyrus for mentalizing 'Theory of Mind' tasks, particularly when tasks involved processing of pictures rather than verbal stories (Gallagher *et al.*, 2000). Since there were no significant group differences bilaterally over the hemispheres, some form of basic mentalizing process might be involved in classification.

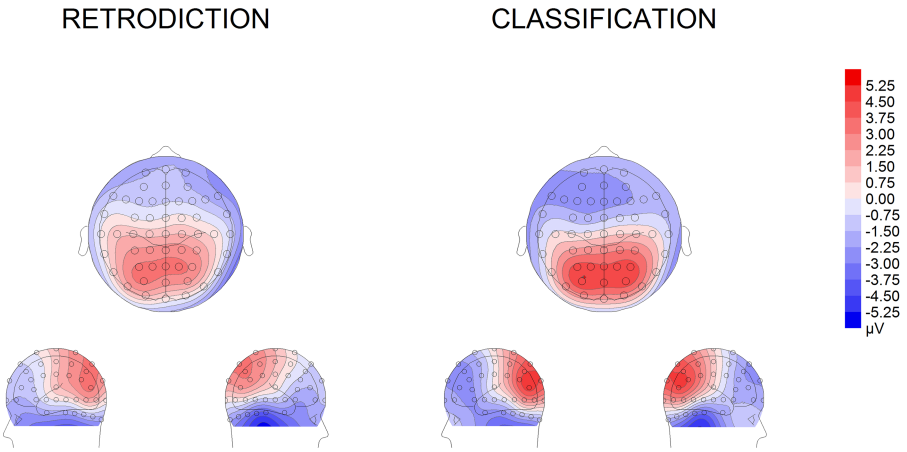

Fig. 7. Voltage maps for mean amplitudes between 800 and 1000 ms, depicting late frontal hemispheric asymmetries for the retrodiction (left) and classification (right) groups. Top view maps use a 110-degree equidistant projection; left and right view maps both use a 90-degree equidistant projection. All maps were obtained using spherical spline interpolation. Electrode positions are also shown. Negativity is blue.

If mentalizing processes were engaged in both tasks, they can be dissociated by the different hemispheric activity at these later components over the frontal and fronto-temporal electrodes. Specifically, elaborated mentalizing processes involved in retrodictive inferences may occur at later components only in the right hemisphere, as indicated by a stronger negativity over right frontal electrodes. In so far as elaborated mentalizing processes were involved uniquely in the retrodiction task, these processes appear to involve right hemispheric frontotemporal areas significantly more than their left hemispheric counterparts.

Based on previous work (Gallese and Goldman, 1998; Teoh *et al.*, 2017), we assume that retrodiction entails the inference of an inner state. Another rather implausible possibility is that neither classification nor retrodiction involves inferences of inner states but merely 'matching' behavior to a face according to a system of behavioural rules (Povinelli and Vonk, 2003; Perner and Ruffman, 2005). Even if perceivers in the current study did not make inferences of inner states, at least the ERP findings suggest that different forms of processing were engaged: perceivers apparently treated the two tasks differently depending on how they were instructed/cued, as demonstrated by the ERP results.

In summary, when classifying an emotional expression from the face, one engages initially in structural face encoding (N170; Bentin *et al.*, 1996; Maurer *et al.*, 2008; Kloth *et al.*, 2010), followed by a preliminary emotional analysis (P200; Paulmann & Pell, 2009) and a cognitive evaluation, which draws upon stored knowledge about the expressions (P300-600; Adolphs, 2002). When one engages in retrodiction, early ERP correlates are not modulated by the different expressions; moreover, emotional decoding and cognitive evaluation of facial expressions are relatively weaker than when classifying facial expressions. Notably, the process involved in retrodiction may be cortically time-consuming, only appearing in later components in the present study.

Supplementary Data

Supplementary data are available at *SCAN* online.

Funding

The study was funded by the Experimental Psychology Society (EPS) Study Visit Grant to K.K.; a Young Researcher Support Grant DRM/2014-02 to D.S. from the Friedrich Schiller University Jena, Germany; a grant by the Deutsche Forschungsgemeinschaft (SCHN 1481/2-1) to D.S.; and a grant by the Deutsche Forschungsgemeinschaft (FOR 1097) to S.R.S.

References

Adams, R.B., Jr., Kleck, R.E. (2003). Perceived gaze direction and the processing of facial displays of emotion. *Psychological Science*, **14**(6), 644–7.

Adolphs, R. (2002). Neural systems for recognizing emotion. *Current Opinion in Neurobiology*, **12**, 169–77.

Aichhorn, M., *et al.* (2006). Do visual perspective tasks need theory of mind? *NeuroImage*, **30**(3), 1059–68.

Apperly, I. (2013). Can theory of mind grow up? Mindreading in adults, and its implications for the development and neuroscience of mindreading. In Simon Baron-Cohen, Michael Lombardo & Helen Tager-Flusberg (eds.), *Understanding Other Minds: Perspectives from Developmental Social Neuroscience*. Oxford: Oxford University Press, 72–92.

Apperly, I.A., Butterfill, S. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological Review*, **116**, 953–70.

Apperly, I.A., Riggs, K.J., Simpson, A., Chiavarino, C., Samson, D. (2006). Is belief reasoning automatic? *Psychological Science*, **17**(10), 841–4.

Ashley, V., Vuilleumier, P., Swick, D. (2004). Time course and specificity of event-related potentials to emotional expressions. *NeuroReport*, **15**, 211–6.

Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., Clubley, E. (2001). The autism-spectrum quotient (AQ): evidence from asperger syndrome/high-functioning autism, males and females, scientists and mathematicians. *Journal of Autism and Developmental Disorders*, **31**, 5–17.

Batty, M., Taylor, M.J. (2003). Early processing of the six basic facial emotional expressions. *Cognitive Brain Research*, **17**(3), 613–20.

Bentin, S., McCarthy, G., Perez, E., Puce, A., Allison, T. (1996). Electrophysiological studies of face perception in humans. *Journal of Cognitive Neuroscience*, **8**, 551–65.

Berg, P., Scherg, M. (1994). A multiple source approach to the correction of eye artifacts. *Electroencephalography and Clinical Neurophysiology*, **90**(3), 229–41.

Blau, V.C., Maurer, U., Tottenham, N., McCandliss, B.D. (2007). The face-specific N170 component is modulated by emotional facial expression. *Behavioral and Brain Functions*, **3**.

Cassidy, S., Mitchell, P., Chapman, P., Ropar, D. (2015). Processing of spontaneous emotional responses in adolescents and adults with autism spectrum disorders: Effect of stimulus type. *Autism Research*, **8**, 534–44.

Cassidy, S., Ropar, D., Mitchell, P., Chapman, P. (2014). Can adults with autism spectrum disorders infer what happened to someone from their emotional response? *Autism Research*, **7**, 112–23.

Cuthbert, B.N., Schupp, H.T., Bradley, M.M., Birbaumer, N., Lang, P.J. (2000). Brain potentials in affective picture processing: covariation with autonomic arousal and affective report. *Biological Psychology*, **52**, 95–111.

Dimberg, U., Thunberg, M., Elmehed, K. (2000). Unconscious facial reactions to emotional facial expressions. *Psychological Science*, **11**(1), 86–9.

Eimer, M. (2000). Event-related brain potentials distinguish processing stages involved in face perception and recognition. *Clinical Neurophysiology*, **118**, 694–705.

Eimer, M., Holmes, A. (2002). An ERP study on the time course of emotional face processing. *NeuroReport*, **13**, 427–31.

Fodor, J.A. (1992). A theory of the child's theory of mind. *Cognition*, **44**(3), 283–96.

Gallagher, H.L., Frith, C.D. (2003). Functional imaging of 'theory of mind'. *Trends in Cognitive Sciences*, **7**, 77–83.

Gallagher, H.L., Happé, F., Brunswick, N., Fletcher, P.C., Frith, U., Frith, C.D. (2000). Reading the mind in cartoons and stories: an fMRI study of 'theory of mind' in verbal and nonverbal tasks. *Neuropsychologia*, **38**(1), 11–21.

Gallese, V., Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences*, **2**, 493–501.

Geangu, E., Gibson, A., Kaduk, K., Reid, V.M. (2012). The neural correlates of passively viewed sequences of true and false beliefs. *Social Cognitive and Affective Neuroscience*, **8**, 423–37.

Kang, K., Anthoney, L., Mitchell, P. (2017). Seven- to 11-year-olds' developing ability to recognize natural facial expressions of basic emotions. *Perception*, **46**, 1077–89.

Kloth, N., Schweinberger, S.R., Kovács, G. (2010). Neural correlates of generic versus gender-specific face adaptation. *Journal of Cognitive Neuroscience*, **22**, 2345–56.

Krombholz, A., Schaefer, F., Boucsein, W. (2007). Modification of N170 by different emotional expression of schematic faces. *Biological Psychology*, **76**, 156–62.

Leslie, A.M. (1987). Pretense and representation: the origins of theory of mind. *Psychological Review*, **94**(4), 412.

Liu, D., Sabbagh, M.A., Gehring, W.J., Wellman, H.M. (2004). Decoupling beliefs from reality in the brain: an ERP study of theory of mind. *NeuroReport*, **15**, 991–5.

Liu, D., Sabbagh, M.A., Gehring, W.J., Wellman, H.M. (2009). Neural correlates of children's theory of mind development. *Child Development*, **80**, 318–26.

Luck, S.J. (2005). *An Introduction to the Event-Related Potential Technique*, Cambridge, London: MIT Press.

Lundqvist, D., Flykt, A., Öhman, A. (1998). *The Karolinska Directed Emotional Faces – KDEF*, Department of Clinical Neuroscience, Psychology section, Karolinska Institutet.

Maurer, U., Rossion, B., McCandliss, B.D. (2008). Category specificity in early perception: face and word N170 responses differ in both lateralization and habituation properties. *Frontiers in Human Neuroscience*, **2**, 18.

McCleery, J., Surtees, A., Graham, K., Richards, J., Apperly, I. (2011). The neural and cognitive time-course of theory of mind. *Journal of Neuroscience*, **31**, 12849–54.

Meinhardt, J., Sodian, B., Thoermer, C., Döhnel, K., Sommer, M. (2011). True- and false-belief reasoning in children and adults: an event-related potential study of theory of mind. *Developmental Cognitive Neuroscience*, **1**, 67–76.

Mitchell, J.P., Banaji, M.R., Macrae, C.N. (2005). General and specific contributions of the medial prefrontal cortex to knowledge about mental states. *NeuroImage*, **28**(4), 757–62.

Orgeta, V. (2010). Effects of age and task difficulty on recognition of facial affect. *The Journals of Gerontology, Series B: Psychological and Social Sciences*, **65**, 323–7.

Paulmann, S., Pell, M.D. (2009). Facial expression decoding as a function of emotional meaning status: ERP evidence. *NeuroReport*, **20**, 1603–8.

Pell, M.D., Monetta, L., Paulmann, S., Kotz, S.A. (2009). Recognizing emotions in a foreign language. *Journal of Nonverbal Behavior*, **33**, 107–20.

Perner, J. (1991). *Understanding the Representational Mind*, London: MIT Press.

Perner, J., Ruffman, T. (2005). Infants' insight into the mind: how deep? *Science*, **308**(5719), 214–6.

Pillai, D., Sheppard, E., Mitchell, P. (2012). Can people guess what happened to others from their reactions? *PloS One*, **7**.

Pillai, D., Sheppard, E., Ropar, D., Marsh, L., Pearson, A., Mitchell, P. (2014). Using other minds as a window onto the world: guessing what happened from clues in behaviour. *Journal of Autism and Developmental Disorders*, **44**, 2430–9.

Premack, D., Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, **1**, 515–26.

Povinelli, D.J., Vonk, J. (2003). Chimpanzee minds: suspiciously human? *Trends in Cognitive Sciences*, **7**(4), 157–60.

Sabbagh, M.A., Moulson, M.C., Harkness, K.L. (2004). Neural correlates of mental state decoding in human adults: an event-related potential study. *Journal of Cognitive Neuroscience*, **16**, 415–26.

Sabbagh, M.A., Taylor, M. (2000). Neural correlates of theory-of-mind reasoning: an event-related potential study. *Psychological Science*, **11**, 46–50.

Sato, W., Yoshikawa, S. (2007). Spontaneous facial mimicry in response to dynamic facial expressions. *Cognition*, **104**(1), 1–18.

Saxe, R. (2007). Theory of mind. In: Ochsner, K.N., Kosslyn, S., editors. *The Oxford Handbook of Cognitive Neuroscience: Volume 2: The Cutting Edges*, Oxford University Press.

Saxe, R. (2009). Theory of mind (neural basis). *Encyclopaedia of Consciousness*, **2**, 401–10.

Saxe, R., Kanwisher, N. (2003). People thinking about thinking people. The role of the temporo-parietal junction in theory of mind. *NeuroImage*, **19**, 1835–42.

Schindler, S., Kissler, J. (2016). People matter: perceived sender identity modulates cerebral processing of socio-emotional language feedback. *NeuroImage*, **134**, 160–9.

Schurz, M., Radua, J., Aichhorn, M., Richlan, F., Perner, J. (2014). Fractionating theory of mind: a meta-analysis of functional brain imaging studies. *Neuroscience & Biobehavioral Reviews*, **42**, 9–34.

Sessa, P., Meconi, F., Han, S. (2014). Double dissociation of neural responses supporting perceptual and cognitive components of social cognition: Evidence from processing of others' pain. *Scientific Reports*, **4**(7424), 1–8.

Sheppard, E., Pillai, D., Wong, G.T.-L., Ropar, D., Mitchell, P. (2016). How easy is it to read the minds of people with autism spectrum disorder? *Journal of Autism and Developmental Disorders*, **46**, 1247–54.

Smith, M.L. (2012). Rapid processing of emotional expressions without conscious awareness. *Cerebral Cortex*, **22**, 1748–60.

Tcherkassof, A., Bollon, T., Dubois, M., Pansu, P., Adam, J. (2007). Facial expressions of emotions: a methodological contribution to the study of spontaneous and dynamic emotional faces. *European Journal of Social Psychology*, **37**, 1325–45.

Teoh, Y., Wallis, E., Stephen, I.D., Mitchell, P. (2017). Seeing the world through other minds: Inferring social context from behaviour. *Cognition*, **159**, 48–60.

Vogeley, K., Bussfeld, P., Newen, A., *et al.* (2001). Mind reading: neural mechanisms of theory of mind and self-perspective. *NeuroImage*, **1**(1), 170–81.

Wagner, H.L. (1993). On measuring performance in category judgment studies of nonverbal behavior. *Journal of Nonverbal Behavior*, **17**, 3–28.

West, T.V., Kenny, D.A. (2011). The truth and bias model of judgment. *Psychological Review*, **118**, 357–78.

Young, L., Dodell-Feder, D., Saxe, R. (2010). What gets the attention of the temporo-parietal junction? An fMRI Investigation of Attention and Theory of Mind. *Neuropsychologia*, **48**(9), 2658–64.